Draft Online Safety (Designated Internet Services and Relevant Electronic Services - Class 1A and Class 1B Material) Industry Standards 2024

Google Submission

23 January 2024

Key points of Google's submission

Google appreciates the opportunity to provide comments on the eSafety Commission's draft Designated Internet Services (**DIS**) and Relevant Electronic Services (**RES**) Standards (together, the **Standards**) to address certain categories of restricted and illegal content online. These are important issues that require thoughtful input from a variety of stakeholders, including online service providers, the Australian Government, civil society and others.

While the Internet has an immensely positive impact on society, we also recognise that there can be a challenging side of open platforms, and that in some cases bad actors have exploited this openness. We understand the sensitivity and importance of these areas and have devoted careful attention to developing an approach in offering our services that limits harm while protecting users' ability to express themselves online. We have not waited for legislation to act in tackling illegal or lawful but potentially harmful content; we have developed our own guidelines and implemented extensive efforts to help prevent and address harmful and unlawful content across our services, including by working appropriately with government, law enforcement, and other stakeholders in Australia and around the world.

Our strategies for tackling illegal and harmful content are tailored to each of our platforms. Across our products, our teams tackle a broad spectrum of online abuse from scams, like the email allegedly from a 'relative' stranded abroad needing a bank transfer to get home safely, to abhorrent content including child sexual abuse material (CSAM) and terror content. Understanding the different parameters of the products we deliver is vital to our work and policy development. Given that breadth, our team is diverse, comprising product specialists, engineers, lawyers, data scientists, ex-law enforcement officials and others. They work hand-in-hand around the world and with a global network of safety and subject matter experts. We now have over 10,000 people across Google working on content moderation and removal across our platforms. This includes reviewers who work around the world, 24/7, speak many different languages and are highly skilled in this task.

For each product, we have a specific set of rules and guidelines that are suitable for the type of platform, how it is used, and the risk of harm associated with it. These approaches range from clear community guidelines, with mechanisms to report content that violates them, to increasingly effective artificial intelligence (AI) and machine learning that can facilitate removal of some types of harmful content before a single human user has been able to access it.

We are supportive of online service providers implementing robust systems to identify and address harmful content online and of regulation that is carefully crafted and appropriately tailored; however we are concerned that some aspects of the current draft could be vulnerable to abuse and may have inadvertent, negative impacts on Australians' right to privacy, security and access to information or services.

We have summarised these concerns below and provide further detail in the body of our submission.

(A) Scope of obligations: requirements to "detect and remove" and "disrupt and deter" known CSAM, CSEM (new and known) and pro-terror material.

- **Where practicable (and unless there are policy reasons to the contrary), the Standards should be consistent with equivalent provisions under the Online Safety Industry Codes (Industry Codes).** In particular, we suggest that the following definitions and/or obligations in the Standards should be aligned with the Industry Codes: (a) definition of "appropriate action" should be consistent with the definition of "appropriate" in the Industry Codes; (b) obligations to "disrupt and deter" should be consistent with equivalent obligations in the Industry Codes; (c) obligations to notify verifying agencies of new CSEM and pro-terror, undertake risk assessments, notify to the eSafety Commissioner of new features, and to report

on compliance with the Standards, should be consistent with (and not more onerous) than equivalent obligations under the Industry Codes.

- **The Standards should make clear that providers are only required to implement policies, systems, processes and technologies to address online harm that are "reasonable" and "proportionate", taking into account all relevant circumstances including other potential adverse impacts on the rights of end-users such as privacy, access to information and information security.** This flexibility can be achieved by amending the definition of "appropriate action" in the Standards to align with the definition in the Industry Codes (see Industry Codes Head Terms, definition of "appropriate" and section 5.1 and 6.1). Sections 20, 21 and 22 of the RES Standard and section 21, 22 and 23 of the DIS Standard should also be amended to clarify that the substantive obligations contained within, only require a provider to take "appropriate action" (in line with the amended definition, as proposed above).

- **Obligations to "detect and remove" known pro-terror material should be limited only to those services where content is widely distributed or shared.** While it is possible to detect with a high degree of confidence known child sexual abuse material (known CSAM), this is not the case with the detection of known pro-terror material. Whether material is "pro-terror" is context-dependent and there are no uniform or globally accepted definitions. It is particularly difficult to accurately detect pro-terror material in closed communication RES or end-user managed hosting services, where the purpose of the content is unclear, and there may be legitimate reasons for users to possess such material (for example, for journalistic, research or other academic purposes).

- **Obligations to "disrupt and deter" child sexual exploitation material (CSEM) and pro-terror material (both new and known) should make clear that providers are not required to deploy technology (such as for instance AI) to detect and identify new CSEM or pro-terror material, particularly for closed communication RES and end-user managed hosting services**. While technology to proactively detect new CSEM, pro-terror and other harmful content is improving every day, it is not infallible and should only be deployed carefully and only where a provider is satisfied appropriate safeguards are in place. We recommend that this section be amended to ensure that providers are able to satisfy this obligation by either implementing systems, processes or by using technology (where appropriate). This is consistent with equivalent provisions in the Industry Codes.

(B) Notification of "pro-terror" material

- **Reporting new "pro-terror material" to an organisation that verifies "pro-terror material" is presently unworkable.** While reporting mechanisms for CSEM are well established, this is not the case for pro-terror material. We are not aware of an organisation that has the capacity, infrastructure or resources to verify in large volumes (in accordance with an Australian specific standard) whether a piece of content is new "pro-terror". This is also not a requirement in any of the Industry Codes. We recommend that this requirement be removed. Participation in, and contribution to industry hash-sharing databases for pro-terror could instead be encouraged as part of a provider's "development plan" or as part of the systems, processes or technology that a provider may deploy to meet the requirements to "disrupt and deter" pro-terror material. This is consistent with the approach taken in the Industry Codes.

(C) Terms of service and enforcing breaches of terms of service or community standards

- **The Standards should ensure that obligations for providers to implement terms of services that prohibit the solicitation, access, distribution of storage of class 1 material (other than CSEM and pro-terror material), and to remove from the service or prevent further access or distribution of class 1 material (other than CSEM and pro-terror material), treat online content consistent with**

offline content, and are proportionate to the risk of harm to other end-users. While some class 1 material is illegal to possess (such as CSEM), this is not the case with all class 1 material (for e.g. certain gratuitous drug related content) and there may be legitimate reasons for the possession, and sharing (to a limited audience) of such content - for example, bystander footage of an extremely violent event (such as a terrorist attack or war crime) emailed to a news organisation. We recommend that section 14 (2) of both Standards be amended to make clear that, in implementing terms of service relevant to class 1 material, providers can have regard to "the purpose and functionality of the service, the legality of the content and the potential harm to other end-users". We also recommend that section 17 (2) (c) and (d) and 25 (2) (c) and (d) of the RES Standard is removed. Instead, providers should only be required to take "appropriate action" (as defined in the Standards - see comments above) in response to a breach of terms of service for class 1 material (other than CSEM or pro-terror) that is proportionate to the extent of the harm to the online safety of Australians that may reasonably be expected to flow from the breach.

- **Requirements for providers to respond to breaches of "community standards" (in addition to breaches of terms of service) are undefined, vague and in our view unnecessary**. To avoid unnecessary complexity, we recommend that references to "community standards" are deleted throughout the Standards.

(D) High risk generative AI DIS and machine learning model platform services

- **New categories for "high risk generative AI DIS" and "machine learning model platform services" would benefit from greater clarity and a broader consultation with industry to ensure the definitions and the minimum compliance obligations are workable.** While we recognise that AI presents challenges and there is a role for regulation to play, we are concerned the proposed definitions are unclear, difficult to apply in practice and the minimum compliance obligations may be unworkable. Challenges with AI are best addressed holistically, as part of the broader review into the Online Safety Act and Basic Online Safety Expectations and the new categories should not at this stage be included for the purpose of the Standards. In our view, it is likely that the services are already caught by one of the other broader categories of services as defined in the DIS Standard, or an Industry Code. If the new categories are included, we still believe it will benefit from greater clarity and broader consultation with industry and welcome the opportunity to discuss further with the eSafety Commission.

(E) Enterprise DIS and RES

- **Definition of Enterprise DIS and RES and reference to "account-holder" is unclear.** Whether a service is an enterprise DIS and RES (or a consumer service) should be defined by the predominant function or purpose of the service, and not the identity of the "account-holder" (or whether the party who is contracting with the provider is an individual or an organisation). We recommend that part (a) of the definition of "enterprise DIS" and "enterprise RES" (which states that "the account holder for which is an organisation (and not an individual)") is deleted.

- **Enterprise DIS should be carved out from any obligation to "deter and disrupt" CSEM or pro-terror material (in line with equivalent provisions in the RES Standard).** Given the nature of enterprise customers, and the contractual and other confidentiality obligations that usually apply to the use of the service, it is unlikely that an enterprise DIS provider would have the capability, visibility, control, or authority to implement effective systems, processes or technologies to "disrupt and deter" end-users from engaging in this type of activity. We recommend that enterprise DIS should not be subject to these obligations, consistent with equivalent provisions for enterprise RES.

(F) Other provisions relevant to both the RES and DIS Standards

- **Provision of information "in-service": Requirements for providers to ensure that information is provided "in-service" may not be practicable or feasible for all service types covered by the Standards**. Instead we suggest that providers should be provided with flexibility as to how best to present and make available certain information to its end-users provided that it does so in a way that is "easily and readily available". This approach is in line with the approach taken in the Industry Codes.

- **Notification of new features and functions to the eSafety Commissioner as soon as a "decision is made" to implement a new feature or function (unless it does not significantly increase the risk that the service will be used to solicit, access, distribute or store class 1A or class 1B material) are vague, onerous and potentially unworkable.** We recommend that obligations to notify the Commissioner of new features or functions align with equivalent obligations under the Industry Codes and be limited only to an obligation to notify of any new feature or function (as implemented) if the service provider reasonably believes it will have a material positive or negative impact on the solicitation, access, distribution or storage of class 1 material.

(G) Other provisions relevant only to the RES Standard

- **Definition of "open communication relevant electronic service (open communication RES): The proposed definition of "open communication RES" is unclear. In particular, the statement that an open communication RES captures a service that "enables an end-user to invite, through use of an internet link, another end-user to communicate with the first end-user" could potentially capture any service that enables a user to share an internet link, including an email address or a link to a video conference.** We understand it is not eSafety's intent to capture services such as email or private video conferencing within the definition of "open communication RES". To avoid confusion, we suggest this statement is removed from the definition.

- **Data retention for "closed communication RES": New requirements for providers to retain user registration information for all end-users of "closed communication RES" and "dating services" (but for no other service) may impose an unnecessary and unacceptable privacy and information security risk to end-users and should be removed.** We are unclear the policy reason why, for closed communication RES (such as email or messaging services) and dating services (but for no other service), user registration information for all users is required to be retained for 2 years. Mandatory requirements for providers to retain user data (after a user has ceased using the service) poses an increased security and privacy risk for all end-users. Where preservation of certain user data is required, we believe that existing and well established processes to obtain preservation orders suffice for this purpose.

(H) Other provisions relevant only to DIS

- **Risk assessment for Tier 1 and Tier 2 DIS: We would be grateful for greater clarity and guidance for how, in practice, a provider is to classify a service as either a Tier 1 or Tier 2 DIS.** As drafted, it is unclear how a provider is to weigh up and assess the different factors identified in section 9 (5) of the DIS Standard to determine whether the risk that class 1A or class 1B material will be generated, distributed or stored on the service is "high", "moderate" or "low". In particular, we are unclear whether in performing the risk assessment, the provider is to assess the risk before or after appropriate mitigations or measures have been implemented. We assume it is the latter, and if so, would recommend this be clarified within the DIS Standard.

(I) Commencement of RES and DIS Standards

- **The commencement period for the Standards should be extended from six months to twelve months. Alternatively, providers should be able to request an additional six month extension where it has reasonable grounds for not being able to meet the initial six month compliance period (similar to the Industry Codes).** Given the variety of different services that will be caught by the RES and DIS Standards, it may be very difficult for providers to meet and ensure compliance with the Standards within six months of registration. We consider twelve months to be more realistic - or alternatively - providers should be provided with an option to seek an extension where it has reasonable grounds for not being able to meet the initial deadline (consistent with the approach taken in the Industry Codes).

Feedback on draft Standards

1. Relationship between Standards and Online Safety Industry Codes ("Industry Codes")

1.1. **We recommend that wherever possible the minimum compliance obligations and definitions used in the Standards should align with the equivalent or similar obligations imposed in the Industry Codes.**

1.2. We appreciate that, for the reasons provided by the eSafety Commission in the discussion paper and fact sheet, the wording, format and some of the terms in the Standards will vary to that in the Industry Codes. We nonetheless consider that wherever possible and in the absence of strong policy reasons to the contrary (for example to address differences in the nature of the services), the nature and scope of the obligations should align with the equivalent obligations under the Industry Codes. This is to ensure a consistent approach is adopted across all online services and for all providers.

1.3. In particular, we note the following key differences between the Standards and the Industry Codes, which in our view should be aligned or consistent for all online services and providers:

- **The definition of "appropriate" and "appropriate action"**. Under the Standards, in determining whether an action taken or proposed in relation to a RES or DIS is "appropriate" the only matter to take into account (other than if it relates to a breach) is if the action "achieves the object of the industry standard in relation to the service". By contrast, "appropriate" is defined under the Industry Codes as measures that are "demonstrably reasonable" in accordance with section 5.1(b) of the Industry Code Head Terms. The test for compliance should be consistent across all online services. We provide further comment on our concerns with regard to the definition used in the Standards in **paragraph 2.1 - 2.8** below.

- **The requirements for providers to "deter and disrupt" CSEM and pro-terror.** The Standards require providers to implement "systems, processes and technologies" that effectively deter and disrupt end-users from using the service to solicit, generate, access, distribute, store or otherwise make available CSEM and pro-terror material. By contrast, the equivalent obligation in the Industry Codes only requires providers to deploy "systems, processes and/or technologies" - that is, a provider is not required to use technology to meet this obligation. We provide further comments on our concerns regarding the approach taken in the Standard in **paragraph 4.1 - 4.8** below.

- **Risk-assessments and reporting obligations**. The requirements to undertake risk assessments and to report on compliance with the Standards are significantly more extensive and onerous than similar obligations under the Industry Codes. For example, the Standards require a provider to report on any risk assessment performed, reports of technical feasibility, reports on outcomes of development programs and compliance reports. The Standards also require that a risk assessment is to be undertaken every time a material change is made to the service (section 8 (5) of both Standards). By contrast, the Industry Codes only require reporting on compliance measures, and a risk-assessment is only required for the purpose of determining the appropriate classification under the applicable Code. It is not clear why more onerous risk-assessment and reporting should be required for RES and DIS providers. In any case, the requirement that a risk assessment must be performed even if it does not change the risk profile or classification of the service under the Standards is redundant - the provider must at any point in time meet the minimum compliance obligations for the service for the applicable service type.

- **Notifying the eSafety Commissioner of new features**. The DIS and RES Standards require providers to notify the eSafety Commission as soon as a decision is made (which may be well before the final feature is developed, implemented and made available to end-users), and at the time a new feature is added to a service, unless the provider determines on reasonable grounds that the change has not significantly increased risks associated with class 1A and class 1B content. By contrast (and while it varies slightly between Industry Codes), the Industry Codes only require notification of new features if a provider considers it likely that the new feature or function will have a material positive or negative effect associated with class 1 material. We provide further comments on our concerns regarding the approach taken by the Standard in **paragraph 9.6 - 9.10** below.

2. Scope of obligations to "detect and remove" known child sexual abuse material (known CSAM) and known pro-terror; and to "disrupt and deter" child sexual exploitation material (CSEM) or pro-terror material (including known CSAM and pro-terror)

[DIS Standard: Sections 21, 22, 23; RES Standard: Sections 20, 21 and 22]

2.1. **Mandatory obligations to implement tools, systems and processes to "detect and remove " known CSAM and known pro-terror, and to "effectively disrupt and deter" the use of the service to solicit, generate, access, distribute, store or make available CSEM (including CSAM) and pro-terror, should only require providers to take steps that are "reasonable", taking in account all relevant circumstances, in line with the approach adopted in the Industry Codes. Providers should be permitted (and in fact encouraged) to ensure that any system, process or technology that they implement is proportionate, effective and does not unreasonably interfere or adversely impact on the rights or interests of the vast majority of end-users.**

2.2. While we are supportive of providers implementing robust systems to identify and address harmful content, we are nonetheless concerned about the potential negative consequences of a broad, inflexible and, potentially unlimited requirement that providers must implement systems, processes and technologies that proactively detect known CSAM and known pro-terror, and

"deter and disrupt" the solicitation, generation, distribution and storage of CSEM and pro-terror material, across an entire service.

2.3. Unlike similar obligations in the Industry Codes, the Standards are not limited by a provider having to take "reasonable" steps (see Industry Codes Head Terms; section 5.1 and section 6). It is also a marked departure from the Basic Online Safety Expectations (**BOSE**), which only require providers to take "reasonable steps". Instead, the only exception to not using a "system, process and technology" is if it is not "technically feasible" for the provider to do so.

2.4. While some (but not all of the obligations) require a provider to take "appropriate action", that definition is still narrower than a test of "reasonableness". While a provider can consider whether a certain action is "reasonable" in response to a breach of applicable terms of service, in all other circumstances (including a decision to deploy proactive detection technology in a file-storage or private messaging service), the only relevant consideration is the extent the action achieves the objective of the industry standard.

2.5. A system, process or technology may be "technically feasible", and may also be effective in "achieving the objectives of the industry standard", but for other legitimate reasons may not be reasonable or appropriate to implement for a particular service. Examples of other relevant and legitimate considerations include:

- **The nature of the service and the impact that the tool, system or process may have on a user's right to, and expectation of privacy**.

  Unlike content that is publicly hosted on a service (for example, on a Tier 1 DIS Service), users have a right to, and an expectation of, privacy in their use of private communication and storage services. A tool, system or process to proactively detect harmful content on a Tier 1 or Tier 2 DIS, may not be suitable for a private communication or file storage service, or may only be suitable for part of the service, under privacy preserving methods, and limited to only certain types of content.

  We acknowledge that in the discussion paper, the eSafety Commission makes clear that the Standards do not require providers to monitor the content of private emails, instant messages, SMS, MMS, online chats and other private communications (see page 6 of the RES Fact Sheet). We are concerned, however, that this is not reflected in the Standards. There is no exception (beyond whether it is technically feasible) that would allow for consideration of whether the tool can be implemented in a privacy preserving manner.

- **The accuracy of the technology to detect harmful content on a service**

  While proactive measures are improving all the time, no technology is infallible and a given technology should only be deployed carefully, and when judged effective by individual companies. Errors in detecting harmful content on services can have a serious and significant impact on the end-user.

  We provide detailed comments on the limitations to detect pro-terror material below. Even the use of hash-matching technology to detect known CSAM images (which is considered one of the most effective and reliable detection methods for such content) is not infallible and still dependent on the quality of the hashes used. This is why, for those services where hash-matching technology is deployed, Google verifies all hashes it

ingests from industry hash-sharing databases before deploying them, and even then, there may be differences in outcomes when matches undergo different human reviews.

We are concerned that unless the Standards provide flexibility to providers in how they design and implement their systems, providers will be required to implement systems without safeguards that adequately protect against unintended and adverse consequences for legitimate users.

- **The adverse impact that a technology, system or process may have on information security.**

  We acknowledge that the eSafety Commission has made clear in its discussion paper that it does not expect companies to design systemic vulnerabilities or weaknesses into end-to-end encrypted (**E2EE**) services, or to do anything that is not technically feasible. While this is welcome, express confirmation within the Standard itself would provide much needed certainty for industry, and would be consistent with the approach taken in the BOSE and the Industry Codes.

  However, while E2EE is one area where concerns arise about the impact a particular tool or technology may have on the security of a user's account, it is not the only area of concern.

  For example, the obligation in the DIS Standard that providers must take "appropriate action" to prevent an offender from re-opening a new account. One means to achieve this would be to collect information from every user to verify their identity (for example, a Government ID) and to store that information indefinitely after a user ceases to use the service. However, this may be contrary to best privacy practice and would impose an unacceptable risk for all users in the event of a security breach. It would also impact other rights users have, such as free access to information and speech.

  While the Australian Government has recognised online security as one of the most serious risks facing Australians, this is not a relevant consideration for providers when assessing whether an action is "appropriate". This is distinct from the Industry Codes, which makes clear that similar obligations do not require a user to "verify an individual's identity or age" (see section 6.1 (f) and (g) of the Industry Codes Head Terms).

2.6. Unless the Standards allow for greater flexibility in how a provider may design its systems, there is real concern that innovation would be stymied and the ultimate design will not strike the best balance for the safety, security and privacy of its users, in line with Google's strong and long-standing commitment.

2.7. We believe this flexibility can be achieved by amending the definition of "appropriate" to align with the definition used in the Industry Codes. The Industry Codes Head Terms define "appropriate" as measures that are "demonstrably reasonable in accordance with section 5.1 (b)". Section 5.1 (b) of the Industry Codes Head Terms clearly set out not only a list of the type of factors that can be taken into account in assessing whether a measure is appropriate, and section 6.1 lists those steps that the Industry Codes do not require (for example, that providers are not expected to design systemic vulnerabilities or weaknesses into E2EE or that providers are not expected or required to monitor private communications).

2.8. While we appreciate that the eSafety Commission may intend to address some of these matters in regulatory guidance, we would suggest that clarification in the Standards would provide much-needed transparency and clarity for the providers who are impacted.

Proposed Amendments

We recommend that:

- The definition of "appropriate action" is amended to align with the definition of "appropriate" under the Industry Codes. The definition should make clear that in assessing whether the use of a certain technology, system or process is appropriate, the provider can take into account the potential adverse impact on the rights of the vast majority of users who do not engage in harmful activities, which would include the type of matters in section 5.1 and section 6 of the Industry Code Head Terms.

- The obligations set out in sections 20, 21 and 22 of the RES Standard and section 21, 22 and 23 of the DIS Standard are amended to clarify that a provider is only required to implement systems, processes and technologies that are "appropriate" (as that definition is amended). For example: "The provider of a service must implement **appropriate** systems, processes or technologies..."

3. Obligations to detect and remove "known pro-terror"

[DIS Standard: Section 22; RES Standard: Section 21]

3.1. **While it is possible to detect with a high degree of confidence known CSAM, this is not the case for the detection of "known pro-terror". Whether material is "pro-terror" is context dependent, there is no uniform or globally accepted definition, and it can be more difficult to accurately detect using technology in private communication or file-storage services. Given existing limitations, we urge eSafety to limit the requirement to detect and remove "known pro-terror" only in those services where content is hosted publicly or is widely distributed or shared.**

3.2. The question of what constitutes pro-terror material varies widely across jurisdictions and is hotly debated as a political matter. While the Global Internet Forum to Counter-Terrorism (**GIFCT**) manages a "hash-sharing" database that includes content contributed by members that meets the organisation's taxonomy, which is publicly available here, each member organisation assesses hashed content in line with their own policies and terms of service. As GIFCT notes on its website:

> "At GIFCT, we respect that each member might operate a little differently. We don't tell our members how to use the hashes or how to apply their own policies."

3.3. Unlike the definition of "known CSAM", the definition of "known pro-terror" (as defined in the Standards) is also not limited to the detection of images. Instead, it may include any material that can be:

> "...detected via hashes, text signals, searches of keywords terms, or URLs or behavioural signals or patterns that signal or are associated with online materials produced by terrorist entities that are on the United Nations Security Council's Consolidated List."

3.4. To comply, providers will require a much broader (and more evasive) array of tools and technologies than would be required to meet the same obligation for the detection and identification of known CSAM. For example, to be able to detect pro-terror material that is associated with "text signals" and "keywords", would presumably require the use of language analysis tools, text classifiers and keyword searches. The deployment of such tools in private communication (or closed communication RES) and file storage services (or end-user managed hosting services) raises significant privacy concerns, including concerns about the monitoring of private communications.

3.5. Whether material is "pro-terror" is also context-dependent. While breakthroughs in machine learning and other technology used to monitor and identify potentially harmful content are impressive, the technology is still evolving and is less accurate for more nuanced or context-dependent content. For example, automated systems that are trained to recognize certain images or patterns of text that may be associated with categories of terrorist content may mistake news coverage, documentaries, educational material, and academic research of these subjects as prohibited content because they contain some of the same images and text.

3.6. Consider a video of a violent attack. In one context, the footage might be documentary evidence of atrocities in areas that journalists have great difficulty accessing. In another context, the footage could be promotional material for an illegal organisation (e.g., a terrorist organisation). And in another, important political speech by marginalised populations. In the same vein, the exact same iconic and horrifying images of historic genocide are used by those who want to advocate for justice and tolerance, on one hand, and those who advocate for violence and further genocide, on the other hand. Between these two poles are those who aspire to report on historic events in an objective manner. Computers cannot yet distinguish this key context. Even a highly trained reviewer could have a hard time telling the difference, and machines are even more limited.

3.7. When it comes to privately stored content, accurately detecting and identifying pro-terror material is even more difficult, given the privacy concerns discussed above and because there are legitimate reasons as to why a user may possess such material - countless journalists, NGOs, and academics use online services to document and study terrorism-related material. For pro-terror material, requirements should take into account how the risk of harm to the public varies between privately stored content relative to broad public dissemination of content. Services such as file storage are not well suited to reaching broad public audiences to disseminate, for example, terrorist propaganda. As such, there is a critical need for a careful and proportionate approach that takes into account differences between services and their use cases.

3.8. Similarly, because possession of pro-terror material may be for a legitimate purpose such as academic study or journalism, mandatory requirements that a provider must remove pro-terror material from the service may not be appropriate or proportionate in all circumstances. Instead, we suggest that the Standard only require that providers take appropriate action or steps to minimise the risk of harm to other end-users caused by the distribution or sharing of that content on the service.

3.9. Given these existing limitations, we urge the eSafety Commission to limit the obligations to "detect and remove" only to that content which is widely shared and distributed, and where the harm, purpose or intent can be much easier to ascertain.

Proposed amendments

We recommend that:

- Mandatory requirements to use technology to proactively detect known pro-terror material should be limited only to the detection of pro-terror material that is publicly hosted or widely distributed and shared.

- Mandatory requirements that providers must remove known pro-terror material from the service should be amended to only require providers to take "appropriate action" (as defined in the Standards - see comments above) to minimise the risk of distribution of the material to other end-users.

4. Obligations for providers to "disrupt and deter" CSEM (including known CSAM) and pro-terror material

[DIS Standard: Section 23; RES Standard: Section 22]

4.1. **While the Standards distinguish between obligations to "detect and remove" known CSAM and known pro-terror material, and to "disrupt and deter" CSEM and pro-terror material, the substantive difference between the two is unclear. We urge the eSafety Commission to clarify that providers are not expected to use technology to proactively identify new CSEM or pro-terror material. Instead, the obligations to "disrupt and deter" can be satisfied by the implementation of either systems or processes and/or the use of the technology. This is in line with the approach taken in the Industry Codes.**

4.2. The substantive difference between obligations to "detect and remove" and "deter and disrupt" is unclear.

4.3. As drafted, the Standards require that providers must "implement systems, processes and technologies" (emphasis added) to effectively deter and disrupt end-users from soliciting, generating, accessing, distributing or storing CSAM or pro-terror material. The examples of tools that could be used to "disrupt and deter" are also those tools that would usually be deployed to "detect and remove" first generation CSEM and pro-terror material. Unlike the obligations to "detect and remove", it is also not subject to whether a system, process or technology is technically feasible. For example, machine learning Classifiers or AI (an example given in the RES and DIS Code) are typically used on services that host publicly accessible content (such as Social Media Services) for the purpose of helping to detect potential new generation of CSEM or pro-terror material.

4.4. We are concerned the Standards could be interpreted as mandating that providers **must** use technologies, such as AI, to proactively detect new CSEM and pro-terror on closed communication RES and end-user managed hosting services. This is a departure from the same obligation in the Industry Codes, which make clear that a provider can choose to implement "systems, processes and/or technologies" (emphasis added) to disrupt and deter the same type of activities - that is, they can rely on either systems, processes or technologies to meet the requirement.

4.5. As noted above, while technology that can proactively detect new CSEM, pro-terror or other harmful content is improving every day, it is not infallible and should be deployed carefully and

only where a provider is satisfied that it has adequate safeguards in place to protect privacy and legitimate or protected speech. While a provider may voluntarily deploy such tools where it considers their use appropriate and it is satisfied that appropriate safeguards are in place, we have concerns about the potential risks of a mandatory requirement for all providers to do so, particularly for closed communication RES (such as email, MMS or SMS) or end-user managed hosting service (such as file-storage services). The nature of these risks is detailed above. Rather than mandate the use of technology (that may not be suitable or appropriate for all services), providers should instead be provided with the flexibility to meet this obligation by deploying one, or a mixture of, systems, processes or technologies.

Proposed amendment

We recommend that obligations to "disrupt and deter" be amended as follows:

> "The provider of a service must implement **appropriate** systems, processes **and/or** technologies that..."

5. Notification of pro-terror material: reporting new pro-terror material to an organisation that verifies "pro-terror" material

[DIS Standard: Section 15; RES Standard: Section 15]

5.1. **We are not aware of an organisation whose role is to receive and verify in large quantities potential new pro-terror material from providers. In the absence of such an organisation, a mandatory reporting obligation is presently unworkable and we recommend that this requirement is removed from the Standard.**

5.2. In the event that CSEM and pro-terror material is identified on a service, section 15 of the RES and DIS Standards impose obligations on providers to: (a) if the material affords evidence of a serious and immediate threat to life or physical safety of a person in Australia, report to enforcement authorities or otherwise as required by law; and (b) if the material is new CSEM or pro-terror material, to report that material to an agency that verifies that material as either CSEM or pro-terror. We understand that, for international companies, the eSafety Commission considers the use of existing reporting channels to be adequate to satisfy these obligations - for example, in the case of CSEM, reporting to NCMEC, but we would welcome further guidance in this regard.

5.3. While mechanisms to report CSEM are well established, we are concerned about the practicality of an obligation to report new pro-terror material to an organisation that "verifies" such material. In particular, it is unclear who and/or which organisation has the capacity and infrastructure to perform such a role.

5.4. An obligation to report new pro-terror material for verification is new - it is not a requirement in any of the Industry Codes (including the SMS Industry Code) and will therefore only apply to a subset of online services. We are also unaware of any similar obligation in any other jurisdiction.

5.5. While Google is a member of the Global Internet Forum to Counter Terrorism (**GIFCT**), which plays an important role in leveraging technology, conducting research and sharing best practices, GIFCT does not perform the role of an independent verifier or recipient of reports of content (similar to or equivalent to the role of NCMEC for CSAM).

5.6. Further, access to and the ability to source or contribute hashes is limited only to members of GIFCT. There are currently approximately 20 members. The majority of providers who will now be subject to the RES and DIS Standard are not members, and may not even be eligible for membership.

5.7. We are also not aware of any organisation in Australia, including the Classification Board, that has the existing resources or infrastructure to, in large volumes, accept, review and verify pro-terror material that has been submitted by providers, and to then make that content available to other providers in the form of a hash-sharing database or engage with law enforcement agencies globally (similar to the existing function of NCMEC for child sexual abuse material). To provide an indication of the potential volume of material that may be reported to an organisation for verification, in the period July 2023 to September 2023, YouTube alone removed 163,417 videos for violation of its Violent Extremism policy (per YouTube Transparency Report - Violent Extremism).

5.8. In the absence of an established independent body or organisation that has the infrastructure or structure to accept and verify pro-terror material at scale (similar to the function or role of NCMEC), a mandatory reporting obligation for pro-terror material is not workable.

5.9. At the same time, the importance of the contribution by providers to programs operated by non-governmental organisations such as GIFCT is well recognised. This is why Google is an active member of GIFCT. In 2021, YouTube contributed over 148,000 hashes to the GIFCT database (see YouTube Transparency Report). As an alternative to a mandatory requirement for all providers to report potential new pro-terror material, the participation in, and the voluntary contribution of new hashes to an independent industry organisation, such as GIFCT, could be encouraged as an example of the type of activity that forms part of a provider's "development plan" or as part of the systems or processes that a provider may implement to comply with the obligation to "disrupt and deter". This approach would be in line with the approach taken in the Social Media Services Industry Code (**SMS Industry Code**) (see for example, section 10 and 16 of the SMS Industry Code).

Proposed amendment:

In the absence of an independent organisation to perform this function, we recommend that section 15(4) of the DIS and RES Standards is deleted. Participation in, and contribution to industry hash-sharing databases for pro-terror material could instead be encouraged as part of the Development Plan or to meet the obligation to "disrupt and deter".

6. Terms of use and responding to breaches of terms of use or "community standards"

[DIS Standard: Sections 14 - 19; RES Standard: Sections 14 - 17]

6.1. **Terms of use - prohibition of the use of the service to solicit, access, distribute or store class 1A material or class 1B material**

[DIS Standard: section 14; RES Standard; section 14]

6.2. **Mandatory obligations for providers to prohibit and remove from the service class 1A and class 1B material (other than CSEM and pro-terror) should be limited only to circumstances where the use, storage or distribution of that material is illegal and/or where it is appropriate or proportionate to the potential harm caused to end-users.**

6.3. Section 14 of the Standards impose broad and mandatory obligations for providers to prohibit (via terms of service) the use of the service to **solicit, access, distribute or store class 1A or class 1B material,** irrespective of whether the possession and/or use of the content is, in all circumstances, illegal. Section 17 (2) (c)and (d) and section 25 (2) (c) and (d) of the RES (but not equivalent provisions in the DIS) also impose mandatory requirements for providers to, in response to a breach of a term of service, remove the material from the service (unless it is not technically feasible) and take steps to ensure that the service no longer permits access to or distribution of the material.

6.4. While it is illegal to possess and access some of the categories of content that falls within class 1A material (for example, child sexual abuse material), it may not be illegal for Australian adults to possess and privately view other class 1 material (for example, certain drug related content which falls within class 1B).

6.5. That the law makes a distinction between the private possession of content that has been Refused Classification (which is not illegal), and its sale, advertisement and distribution (which is illegal), is deliberate: it is to limit the unreasonable intrusion into the private lives of its citizens, particularly in circumstances where there is no identifiable harm to other members of society.

6.6. Similarly, there may be legitimate (or non-malicious) reasons why a user may possess class 1 material (other than CSEM or pro-terror) and may share that material to a limited audience using a relevant electronic service (such as an email, MMS or SMS). For example - bystander footage taken on a user's device of an extremely violent event (for example a terrorist attack or a war crime), uploaded to a user's personal end-user managed hosting service and emailed to a news organisation.

6.7. We urge the eSafety Commission to ensure the Standards treat online content the same as offline content. Where the Australian Government believes a category of content is sufficiently harmful such that even the private possession of that content should be prohibited, the Australian Government may make that content illegal, through transparent and democratic processes and in a necessary and proportionate manner. It should not be done indirectly via Industry Standards and only applicable to online content.

Proposed amendment

We recommend that:

- The following minor amendment is made to section 14(2) of both Standards:

> "...impose an obligation on the account holder of the service to ensure that the service is not used...to solicit, access, generate, distribute or store (**as applicable, having regard to the purpose and functionality**

of the service, the legality of the content and the potential harm to other end-users) class 1A material and class 1B material."

- Mandatory requirements that RES providers must "remove" and "prevent access or distribution" of class 1A and class 1B material (other than CSEM or pro-terror) are deleted from section 17 (2) (c)and (d) and 25 (2) (c)and (d) of the RES Standard. Instead, providers should only be required to take "appropriate action" (as defined in the Standard - see comments above) in response to a breach relating to class 1 material (other than CSEM or pro-terror) that is proportional to the extent of the harm to the online safety of Australians that may reasonably be expected to flow from the breach.

6.8. **Responding to breaches of "community standards"**

6.9. **Requirements for providers to respond to breaches of "community standards" (in addition to breaches of terms of service) are undefined, vague and unnecessary. In our view, reference to "community standards" is not required and to avoid unnecessary complexity, recommend it be deleted throughout.**

6.10. We would welcome greater clarity as to the intended purpose and meaning of "community standards" as used throughout the Standards. While providers should have in place processes to enforce their Terms of Service, the additional requirement that providers must also take appropriate action in response to a breach of "community standards" is not defined in the Standards.

6.11. We are also unclear as to the reason for its inclusion. The Standards require providers to have in place terms of service that address class 1 material, and to take appropriate action to enforce those terms of service. In our view, an additional requirement that providers must also take action in response to a breach of "community standards" for the same type of content is duplicative, unnecessary and risks creating confusion.

Proposed amendment:

We recommend that reference to, or the concept of, a 'community standard' be removed throughout both Standards.

6.12. **Responding to breaches of terms of use or community standards - section 17 (2) (d) (ii) of the DIS Standard**

6.13. **A provider's required response to a breach of terms of service in connection with the distribution, solicitation or storage of CSEM or pro-terror material should not differ depending on whether a user is an adult or under the age of 18 (but still above the age permitted to use the service). In both circumstances, a provider should only be required to terminate an account where it has been used with the intention of causing harm or in response to repeated breaches of terms of service. We recommend that section 17(2)(d)(ii) of the DIS Standard (which does not appear in the equivalent provision for the RES Standard) is deleted.**

6.14. In response to a breach of terms of service in connection with the distribution of CSEM or pro-terror material to end-users, section 17(2)(d)(ii) of the DIS Standard imposes a different (and

stricter) response that a provider must take depending on the apparent age of the end-user - for an adult user, a provider is only required to terminate the account in the event content is distributed to cause harm, or where the end-user has repeatedly breached terms of service; on the other hand, if the user is a child (under 18 years), the provider must terminate the account irrespective of whether the breach was intentional or malicious.

6.15. This section does not appear in the RES Standard, nor any of the Industry Codes. While the SMS Industry Code has a similar obligation, it is limited only to taking action when the user is below the age permitted by the service's terms of service (typically 13). We are unclear why a different approach is required for DIS services, and/or whether this is intending to refer to a child under the age permitted by a provider's terms of service

6.16. Even where that user may be under 18, there may still be circumstances where the termination of a user's account is not reasonable or proportionate in response to the detection of CSEM or pro-terror material. For example, if the surrounding circumstances suggest that the sharing, distribution or storage was not malicious or intentional. For the purpose of this section, in any decision as to whether an account should be terminated, we consider the same set of factors should apply irrespective of the user's age (provided that the user is otherwise over the age permitted by the terms of service to hold an account).

Proposed amendment:

We recommend that section 17 (2)(d)(ii) of the DIS Standard is deleted.

7. Enterprise DIS and RES

7.1. **Definition of enterprise DIS and RES**

7.2. Enterprise DIS and RES are defined as services limited only to where the account-holder (defined as the person who is the counterparty to the agreement for the provision of the service) is an organisation and not an individual. While we agree that "enterprise DIS" and "enterprise RES" should be directed at, or limited to, those services where the primary functionality function is to enable an account-holder to use the service for a commercial or enterprise purpose (as opposed to personal use), the identity of the "account holder" is not determinative of the purpose for which a service is being used.

7.3. Depending on the size and structure of a business (for example, whether the business is a partnership or a sole-trader), there may be many reasons why an account-holder may be an individual rather than the business itself, and in many instances, it will not be practical or feasible for a provider to distinguish between different account-holders, for the same service, based on the identity of the contracting party. For that reason, we recommend the deletion of part (a) from the definition of "enterprise DIS" and "enterprise RES"

Proposed amendment

We recommend that part (a) of the definition of "enterprise DIS" and "enterprise RES" (which states that "the account holder for which is an organisation (and not an individual)") is deleted.

7.4. Minimum compliance obligations for enterprise DIS

[DIS Standard: Section 23]

7.5. The DIS (but not the RES) Standards require enterprise DIS to implement systems, processes and technologies to effectively deter and disrupt end users from soliciting, generating, accessing, distributing or storing CSEM or pro-terror material. Given the nature of enterprise services, "enterprise DIS" should not be subject to any obligation to "disrupt and deter" CSEM and pro-terror material (in line with equivalent obligations for enterprise RES).

7.6. Enterprise DIS (but not enterprise RES) are subject to broad obligations to implement systems, processes and technologies that "effectively deter end-users of the service from using the service" and "effectively disrupt attempt[s] by end-users of the service" to solicit, generate, access, distribute, store or otherwise make available CSEM and pro-terror material. Additional obligations also apply to "enterprise DIS" which provides pre-trained machine learning models for integration.

7.7. There is no clear policy reason why enterprise DIS (but not enterprise RES) is subject to an obligation to "disrupt and deter" the solicitation, generation, access, distribution or storage of CSEM and pro-terror material. In any case, given the nature of enterprise customers, and the contractual and other confidentiality obligations that apply to the use of the service, it is unlikely that an enterprise DIS provider would have the capability, visibility, control, or authority to implement effective systems, processes or technologies to disrupt and deter end-users from engaging in this type of activity.

7.8. Consistent with the approach taken for enterprise RES, we recommend that enterprise DIS be removed from section 23 and from any obligation to "disrupt and deter" CSEM and pro-terror material.

Proposed amendment

We recommend that section 23 (1) (b) ("enterprise DIS") is deleted from the DIS Standard.

8. High risk generative AI DIS, machine learning model platform service and enterprise DIS (with a machine learning platforms service)

8.1. **While we recognise that Artificial Intelligence (AI) presents challenges and there is a need for regulation, we consider this is best done holistically and as part of the broader review currently being undertaken by the Australian Government and/or as part of the broader review into the Online Safety Act and Basic Online Safety Expectations. If the eSafety Commission proceeds with the new categories of service, we still consider that the definitions and minimum compliance obligations would benefit from greater clarity and a broader consultation with industry. As drafted, we are concerned that they are unclear, difficult to apply in practice and the proposed minimum compliance obligation may not be appropriate or suitable for all new or existing uses of this technology.**

8.2. The DIS Standard includes new, specific categories of service for “high risk generative AI DIS" and “machine learning model platform service”. Neither appeared, nor were contemplated as part of the draft DIS and RES Industry Codes and have not been subject to prior industry consultation.

8.3. We recognise that, while Artificial Intelligence (AI) has the potential to unlock significant economic, scientific and social progress for nations like Australia and the wider global community (including in the context of online safety), it presents challenges and must be developed with utmost care and responsibility.

8.4. For our part, we already place a heavy emphasis on child safety when developing our own generative AI products. This includes:

- following Google’s responsible Generative AI Principles in protecting all of Google’s publically available models and services built on top of those models;
- working with other stakeholders in the child safety ecosystem to share and understand best practices, including chairing the Generative AI and CSAM Working Group for the Technology Coalition, and participating in working groups with Thorn in developing best practices in this field; and
- deploying a variety of child safety and CSAM protections in our models, including in training data and at the level of user prompts.

8.5. We are also supportive of regulation for AI that is risk-based and proportionate and which, while mitigating potential risks, still preserves a pro-innovation regulatory environment. To that end there are already a number of separate Australian Government initiatives underway to investigate and consult on how best to ensure the safe and responsible use of AI in Australia (see for example the Australian Government’s interim response to the safe and responsible AI Consultation) including a separate and targeted consultation as part of the broader review into the Online Safety Act and Basic Online Safety Expectations. In light of these initiatives, we suggest that rather than adopt a piecemeal approach to AI regulation, it may be preferable for the eSafety Commission to consider the challenges posed by AI holistically as part of these broader consultation processes.

8.6. If, however, the new categories are to be included in the current Standards, we still consider the proposed provisions would benefit from greater clarity and a broader consultation with industry to ensure the definition and minimum compliance obligations properly capture the services and/or address the identified risk, without unnecessarily or adversely restricting innovation in this area. We address these concerns in detail below and welcome the opportunity to engage with the eSafety Commission further on this issue.

- **Definitions of “high risk generative AI DIS” and “machine learning model platform service”**

8.7. Based on the discussion guide, we understand the intent is for machine learning platform model services to be limited only to “open source” or “AI libraries" whereas a high risk generative AI DIS is intended to be services that are provided directly by the provider to an end-user or Consumer. For all other services that may use generative AI technology, these are to be treated consistent with the service that is its predominant function - for example, an email service that may use or incorporate generative AI will still be classified as a “closed communication RES”.

8.8. While this may be the intent, we are concerned it does not match the definitions used in the Standard.

8.9. For example, a “machine learning model platform service” is defined as a DIS with the “predominate functionality of making available one or more machine learning models and making those models available for download” whereas a “high impact generative AI DIS” means a DIS that “uses machine learning models to enable an end-user to produce material and for which it is reasonably foreseeable risk that the service could be used to generate synthetic high impact material”. On the other hand, an enterprise DIS which provides “pre-trained machine learning models for integration into a service deployed by an enterprise customer” is to be treated as a separate category to both a machine learning model platform service and an high impact generative AI DIS.

8.10. Given that an enterprise DIS and a high impact generative AI DIS may include machine learning model platform services that could be made available for download by an end-user, and that an enterprise DIS may enable an end-user to produce high impact material, or that a machine learning model platform service may only be provided to enterprise customer, a service could potentially fall within all three categories.

8.11. Additionally, the intended scope of “high impact generative AI DIS” service is unclear. While high impact generative AI DIS is defined as a service that has a “reasonably foreseeable risk that the service could be used to generate synthetic high impact material”, it is unclear whether this is to be assessed before mitigations and guardrails are put in place, or after. We understand it is the eSafety Commission's intent to only capture the latter, in which case, we suggest adding the words:

> “...the service **provided to the end-user** could be used to generate synthetic high impact material”.

8.12. Further, services that would be caught by these new definitions are likely to be already caught by another similar or broader category of service. For example, a high risk generative AI DIS would fall into the category of either a High Impact DIS or a Tier One DIS. For a machine learning model platform service, if the service is the provision of a particular model for use by a developer, that service may either be a Tier 1, 2 or 3 DIS (or if it is only provided to enterprise customers, an enterprise DIS) depending on the nature and risk posed by the model. On the other hand, if the service is merely hosting various models for purchase or download, then that service may fall within the category of a Third Party Hosting Service. Or, if a service utilises one or more machine learning model platform services in its creation, the end-service itself would fall within an existing category of service.

8.13. Rather than introduce new categories of services (which risks creating confusion and complexity for providers in how to apply the Standards), consideration should first be given to whether these services are already adequately caught or addressed by existing categories. If the new categories are required, greater clarity is needed as to how the definitions are to be applied and/or as to which types of services are intended to be caught.

- **Minimum compliance obligations for high risk generative AI DIS, Machine learning model platform services and enterprise DIS (which provide machine learning models for integration into a service)**

8.14. While we recognise there may be a role for regulation to play in respect of generative AI, we are concerned that the proposed minimum compliance obligations may not be suitable or appropriate for all the potential services caught by these new categories. For example:

(a) While the intent is for machine learning platform service to be limited only to "open source" AI Libraries (or similar types of services), the Standard requires providers of machine learning platform service to (among other obligations):

- report CSEM and pro-terror material on its services;
- respond to and remove instances of CSEM or pro-terror material from its service; and
- provide information about eSafety to end-users and provide a mechanism for end-users to report, and make complaints about information on its services (in-service).

However, given the nature of such services, a provider may not have sufficient control, visibility or access to give effect to all of these obligations. In many instances, the person best placed to ensure compliance is the developer who makes available the end-product that is based on or using one or more foundational models (and if that end product is made available to Australian users, that product or service would itself be subject to the Online Safety Act).

(b) Similarly, section 23 (4) imposes an additional obligation on enterprise DIS to "implement and use systems, processes and technologies that automatically detect and flag CSAM in training data". It is not clear if this is referring to the provider's training data or the enterprise customers training data; or whether the intention is instead for the provider to make available to Enterprise Customers settings or guardrails that an enterprise customer could use to help detect CSAM. For reasons already stated above, depending on what the enterprise DIS is, given the nature of enterprise customers, and the contractual and other confidentiality obligations that usually apply to the use of the service, a provider may not have the capability, visibility, control, or authority to monitor a customer's training data.

Proposed amendments

We recommend that:

- New categories for high risk generative AI DIS and machine learning model platform service should not at this stage be included as separate categories of services in the DIS Standard. While regulation may have a role in protecting against the potential harm caused by generative AI technology, this may be best addressed holistically, as part of a broader consultation with Industry. We suggest this be undertaken as part of the forthcoming review of the Online Safety Act and Basic Online Safety Expectations.

- If the new categories are included, we encourage the eSafety Commission to provide greater clarity as to the services that are intended to be caught, and to engage in a broader consultation with industry to ensure the minimum compliance obligations proposed are appropriate, proportionate and achieve the intended objective.

9. Other provisions relevant to both RES and DIS Standard

9.1. **Requirements to provide information to end-users about safety tools, reporting of content, the role of the eSafety Commission and a complaint mechanism "in service"**

[DIS Standard: Sections 28, 29, 32; RES Standard: Sections 26, 27, 28, 31, 32]

9.2. **Requirements that certain information about tools, settings and safety information must be provided "in-service" may not be practical for all types of services - for example, apps or private messaging services such as SMS or MMS. Rather than require this information to be provided "in service", providers should have flexibility as to how best to present this information, provided the information is presented in a way that is "easily and readily accessible" to end-users. This is consistent with the approach adopted in the Industry Codes.**

9.3. In various sections, the Standards require certain information about tools, settings and other safety information (including information about how to report content to the eSafety Commission or to complain to providers about non-compliance with a Standard) to be provided "in-service". These obligations appear to proceed on the assumption that the services caught by the Standards are all websites and Australian specific.

9.4. While this may be feasible for services that are websites, it may not be practical for other types of services such as apps or messaging services. For these types of services, detailed information about a service is often provided in a centralised location such as a Help Center or Safety Center, which can be readily updated by providers and is often more readily and easily accessible by end-users. This is particularly the case for providers (such as Google) that provide multiple services that are often used in conjunction with each other, and where the applicable tools and settings may overlap (for example, the use of Google Family Link for Google Accounts).

9.5. Rather than require such information to be provided "in service", providers should be allowed flexibility with how and where they choose to present this information, provided that this information is presented in a way that is "readily and easily accessible" to end-users. This is consistent with the approach adopted in the Industry Codes.

Proposed amendment

We recommend that obligations to provide certain information about tools, settings and online safety "in-service" be removed. Instead, providers should be required to ensure that such information is presented in a way that is "readily and easily accessible" by end-users.

9.6. **Requirements to to notify the eSafety Commissioner of new features**

[DIS Standard: Section 36; RES Standard: Section 35]

9.7. **The requirement that providers must notify the eSafety Commissioner as soon as practicable after a decision has been made to add a new feature or function (unless it does not significantly increase the risk that the service will be used to solicit, access, distribute or store class 1A or class 1B material) is vague, onerous and potentially unworkable. We**

recommend that obligations to notify the Commissioner of new features or functions align with equivalent obligations under the Industry Codes.

9.8. The Standards require providers to notify the eSafety Commissioner as soon as reasonably practicable after a decision has been made to add a new feature or function and also once that feature has been added to a service, unless the provider determines on reasonable grounds that the change has not significantly increased the risk that the service will be used to solicit, access, distribute or store class 1A or class 1B material.

9.9. The requirement that providers must make such notification "as soon as reasonably practicable" after a decision is made to implement a feature or function is vague, onerous and will in practice, be difficult to comply with. There are many decisions that occur between the initial decision to implement a particular feature, and the final function or feature that is implemented (which may be very different from that originally contemplated). A provider may therefore be required to make repeated and regular reports during the development phase of a particular feature as that feature evolves, and original decisions are superseded. While this will be extremely onerous for providers, we assume the only information relevant to the eSafety Commission's function will be information about the feature or function that is ultimately made available to end-users.

9.10. To ensure these obligations are workable, we recommend that the obligation instead align with equivalent obligation under the Industry Codes. For example, the equivalent obligation under the SMS Industry Code (section 19) requires providers to take "reasonable steps to ensure that eSafety Commission receives updates regarding significant changes to the functionality of their service and that are likely to have a material negative effect on the access or exposure to, or distribution of class 1A and class 1B material". It also makes clear that providers may choose to provide this information in its annual report to eSafety and that the provider is not required to disclose information that is confidential.

Proposed amendment

We recommend that section 35 be amended to align with equivalent obligations under the Industry Codes. In particular, the obligation to provide information to the eSafety Commissioner should be limited only to when a new features or functions (as implemented and after all mitigations have been put in place) poses a material increase in risk to end-users of the solicitation, distribution or storage of class 1A or class 1B content. It should also make clear that providers are not required to disclose confidential information.

10. Provisions specific to DIS Standards

10.1. **Risk assessment for Tier 1 and Tier 2 DIS Services**

[DIS Standard: Sections 8, 9 and10]

10.2. **Greater clarity and guidance is needed as to how, in practice, a provider is to classify a service as either a Tier 1 or Tier 2 DIS.**

10.3. As drafted, it is unclear how a provider is to weigh up and assess the different factors identified in section 9 (5) to determine whether the risk that class 1A or class 1B material will be generated, distributed or stored on the service is "high", "moderate" or "low".

10.4. For example, any service that allows end-users to post content (which would include sites that allow users to leave reviews) poses a risk that class 1A or class 1B material will be accessed, generated or distributed or stored on the service; however, as a result of mitigations in place (including technology to filter and remove offensive or inappropriate content) the actual risk that an end-user will encounter such content on that service may be low. We are unclear whether in performing the risk assessment, the provider is to assess the risk before or after appropriate mitigations or measures have been implemented. We assume it is the latter and if so would recommend this be clarified within the Standard.

10.5. Unless this clarity is provided, we are concerned services may be inadvertently caught by the definition of Tier 1 DIS, and subject to obligations that are not proportionate to the actual risk posed by the service, including, for example, restrictions preventing users under the age of 18 from accessing the service.

Proposed amendments

We recommend that the DIS Standard be amended to clarify that in undertaking the risk assessment, the provider is to assess the risk that the Service, as it is provided to the end-user (and after all appropriate mitigations and controls have been put in place), poses a high, moderate or low risk of an end-user having access to, or being able to distribute, class 1A or class 1B material on the Service.

11. Provisions specific to RES Standards

11.1. **Definition of "open communication RES"**

[RES Standard: Section 6]

11.2. **The definition of "open communication relevant electronic service" has been amended from the definition used in the draft RES Code to include the qualifying statement that it extends to a service that "enables an end-user to invite, through use of an internet link, another end-user to communicate with the first end-user". The distinction between an "open communication relevant electronic service" (open communication RES) and a "closed communication system relevant electronic service" (closed communication RES) should not depend on whether the service enables an end-user to invite, through the use of an internet link, another end-user to communicate; rather an open communication RES should be limited only to those services that enable an end-user to search for and contact a stranger via that service**.

11.3. It is not clear why the sentence "a relevant electronic service that enables an end-user to invite, through the use of an internet link, another end-user to communicate with the first end-user" has been included in the definition of open communication RES.

11.4. As drafted, this could potentially capture any communication service where a user's contact information can be shared by an electronic link, even where that link can only be provided to an

end-user that is already known. For example, the sharing of an email address could be viewed as an invitation to communicate via an internet link; similarly, the sending of a link to access a private video-conferencing tool may also meet this definition.

11.5. The key distinction between a closed communication RES and an open communication RES should be whether an end-user is able to search for and contact another end-user (without previously knowing or having the contact details of that end-user) via the service. It should not depend on whether an end-user can invite another end-user to communicate via an internet link, which may otherwise be private and only able to be shared between end-users who are known to each other. In our view, the inclusion of this statement is confusing, may capture unintended services and is unnecessary.

Proposed amendment

We recommend deleting the following sentence from the definition of "open communication service":

"To avoid doubt, it includes a relevant electronic service that enables an end-user to invite, through use of an electronic link, another end-user to communicate with the first end-user"

11.6. **Data retention requirements for "closed communication RES"**

[RES Standard: Section 19 (8)]

11.7. **Section 19 (8) of the RES Standard requires that for a closed communication RES, providers are to retain registration information (including for example, phone number, email address or other identifier) for a period of 2 years. The same data retention requirements do not apply to any other online service and may pose increased security and privacy risks for all Australian end-users. We recommend that this section is removed.**

11.8. The requirement to retain user information for all end-users after a user may have ceased using the service is not a requirement for any other service that falls under the RES (with the exception of dating services) or DIS Standard, nor any service covered by an Industry Code. We are unclear the policy reasons why a different approach is required for only these two services.

11.9. In any case, the retention of information about a user after a user has ceased using the service, irrespective of whether there is any evidence that a user has breached or engaged in any harmful activity, may pose potential security and privacy risk for Australian users. There are already well established processes for authorities to request preservation orders in appropriate circumstances. We also note the Australian Government recently announced it will conduct a review of existing mandatory data retention requirements (in particular those imposed on Telecommunication providers) following recent large-scale cyber-attacks in Australia as part of the 2023-2030 Cyber Security Strategy.

Proposed amendment

We recommend that section 19(8) is deleted.

12. Timing of commencement of the Standard and release of supporting regulatory guidance

[DIS Standard: Section 2; RES Standard: Section 2]

12.1. **We suggest the commencement period for the Standards (section 2) should be extended to twelve months after the day the Standard is registered, or alternatively, a provision should be made that allows for providers to request an additional six month extension where there are reasonable grounds for not being able to meet the initial six month compliance period. Supporting regulatory guidance should also be released for consultation or review by Industry well before registration of the Standard.**

12.2. Unlike the Industry Codes, which provided an allowance for providers to request an additional six month period to implement specific measures where there were reasonable grounds for not being able to meet the six month compliance period, there is no allowance for providers to seek a similar extension under the Standards.

12.3. Given the significant departure of the Standards from the draft DIS and RES Codes, a six month deadline for full compliance with the Standards will be significantly difficult for many providers to meet (particularly where significant engineering work is required) and that a twelve month transition period is more realistic. We recommend that the transition period be adjusted to either twelve months from commencement of the Standard, or alternatively, that a provision be included that allows a provider to request an additional six month extension period where it has reasonable grounds for not being able to meet the initial six month deadline.

12.4. Similarly, we understand that the eSafety Commission intends to provide supporting regulatory guidance to industry before or after the registration of the Standards. Given the breadth of many of the obligations, it is critical that industry is provided with this guidance well before the registration of the Standards (and in fact, given the technical nature of many of the obligations, be given an opportunity to consult with the eSafety Commission on the form and content of the guidance) to ensure industry understands and can properly assess what is required to comply.

Proposed amendment

We recommend that:

- Section 2 (Commencement) be amended to twelve months, or alternatively, allow for providers to seek an additional six months extension where it has reasonable grounds for not being able to meet the initial six month deadline.
- Draft regulatory guidance is provided to Industry for consultation well before the commencement of the Standard.

Conclusion

We thank the eSafety Commission for the opportunity to review and comment on the Draft Standards, and remain available to provide further information and answer any questions on these materials as required.

[Ends]